Sleepy Dwarf's Somniloquy on "Drowsy Logic" Chip Design

Giovanni Lostumbo
Imagineer Department eV Renewable
Photons

Walt Elias Disney Chair Professor Emeritus
National Imagination Laboratory

+1(708) 303-8175

giovanni.los@proton.me

Abstract

Power-efficiency has been an increasing design consideration in virtually all new silicon in the past 15 years. Power-first[1] designs, however, typically appear only in niche applications such as IoT. A 2023 retrospective paper describing a research lab's 2002 circuit, using a technique called "drowsy logic," reviewed historical strategies to limit leakage in the context of foundries' recent implementation of low-leakage FinFET and Gate-All-Around technologies.[2,3] This review explores new research and additional industry applications of drowsy logic.

CCS Concepts

• General→ • System Architecture→Power efficiency

B.3.2 [Cache Memories]: Design optimization for low power

Keywords

Cache memory, Low power, Gate leakage

1. Introduction

In *Computer Architecture Techniques for Power-Efficiency* (2008), drowsy circuits are defined as "a new class of state-preserving leakage reduction techniques."[4a] When factored into the dynamic power equation $P = CV^2 A f$, it shows that there can be a cubic reduction in power without a proportional reduction in performance,[4b] especially in memory-bound or latency-tolerant regions of code.[4c] When combined with sub-threshold voltage, it allows static power to be greatly reduced.[4d,4e] The original 2002 technique involves a strategy, called the "simple" policy, of placing all lines in a drowsy mode using a single global counter, awakening it only when it is accessed.[3] The performance trade off was known to reduce leakage up to 85% while increasing run-time by just 0.62% in certain conditions (using 93% drowsy lines). The paper focused on advanced drowsy strategies in reducing latency due to L1's time-critical cache, but suggested L2 strategies could use the simpler techniques.

2. A Comparison to Previous Techniques

2.1 Vdd gating, Cache Decay, and Adaptive Mode-Control

The 2002 paper and the 2023 retrospective paper contrasts drowsy logic with three previous techniques to reduce leakage developed in 2000 and 2001. The first, called *gated-Vdd*,[5] used a circuit-level technique to gate supply voltage to reduce leakage in unused SRAM. That technique, paired with a novel resizable cache architecture (DRI i-cache), was said to reduce leakage by 62% with a minimal impact on performance. The second, Cache Decay, used a time-based strategy to turn off a cache line after a pre-set number of cycles have elapsed since its last access.[6] This method is said to reduce L1 leakage up to 70% using competitive algorithms. The third, Adaptive Mode Control (AMC), used tags that are always active, tracks which cache lines are missed, and can adjust the number of intervals to turn off cache lines based on previous misses.[7] Cache Decay was also recently reviewed by their authors in the 2023 ISCA@50 Retrospective.[8]

2.2 A Systems-Level Design

A 2008 paper titled "BTB Access Filtering: A Low Energy and High Performance Design" describes lowering branch target buffers and using direct-mapped BTBs in superscalar processors with drowsy techniques in the filter buffer to limit predictor energy consumption by 92.7%, with up to a 10.8% performance trade-off[9]. Early ARM processors, such as ARM7 and ARM9, did not use superscalar architecture and likely did not need large buffers.[10] Static branch prediction was used, however, in power-conscious processors such as the ARM810.[11,12] Significant research has been explored in developing more power-efficient superscalar architecture.[4f]

2.3 Direct-Mapping

Direct-mapping of hardware registers appears to have a similarity to MMU-less operating systems, such as μClinux[13], developed by Jeff Dionne and Ken Albanowski in 1998, which used flat memory addresses and was further developed by companies such as EmCraft.[14] In software, Cortex-M processors feature interrupt/exception handling, where instead of automatically putting floating point registers onto the stack, can be configured to do so in a "lazy" way.[15-17] While most processors are designed to be fast, or to reduce latency, they also benefit from using tricks like that to make them less resource intensive. Direct mapping of both hardware and kernel resources offers a path to limit energy-intensive memory management caches. As a chip can be designed to operate in a more deterministic manner as in real-time

operating systems (RTOS), use cases involving known application resource limits and scheduling can be increasingly factored into software-defined hardware (SDH), a decades old design goal.[18,19] One design concept that monolithic chips use is "holistic timing."[20] This is where multiple systems fit on a single die and operate within a certain window, including breaking partitions in the clock. A Power-First design would most certainly benefit from opting towards fewer Die-to-die interfaces, which typically increase power consumption. An advantage to using chiplets, such as Bunch of Wires, however, would be the parallel or co-design and integration of peripherals such as lower power radios and I/O by third parties.[21] While sub and near-threshold processors are a major focus of power reduction, memory can occupy 2-8x+ the size of a single microcontroller such as Zero-riscy (now Ibex) or RI5CY (cv32e40p), according to PULP Platform of ETH Zürich.[22] General purpose operating systems require far more memory and cache than microcontrollers. Efficient design of both hardware and software using the above techniques can lessen the amount of memory (and thus power) needed.

3. Drowsy logic in modern process nodes

Industry news articles appear to have decreased mentions of "drowsy" circuits in the mid 2010s. A keyword search of one well-known site turned up only two mentions in 2014 and 2015 for "drowsy:" "For memories, people are building additional operational modes for them, such as drowsy modes."[23,24] A review in the ICSA@50 retrospective paper states drowsy logic is "a form of voltage scaling": "We proposed a design in which one can choose between two different supply voltages in each cache line, corresponding to normal supply voltage and a drowsy lower voltage. In effect we used a form of voltage scaling to reduce static power consumption."[2] While the 2002 paper cites voltage scaling precedents, its novel feature was that it was applied to static power: "Such a dynamic voltage scaling or selection (DVS) technique has been used in the past to trade off dynamic power consumption and performance. In this case, however, we exploit voltage scaling to reduce static power consumption."[3] In other words, the term became part of DVS from the start. New techniques utilizing hybrid or novel implementations of drowsy transistors in SRAM continue to be researched in academic labs.[25-27] Techniques sometimes have multiple industry names. As early as 1992, multi and dual-threshold voltage techniques have been described.[28,29] Drowsy logic still offers some of the best energy savings: "Moreover, since the penalty for waking up a drowsy line is relatively small (it requires little energy and only 1 or 2 cycles) and there are less frequent accesses to the lower memory hierarchy than Gated-V_{DD} schemes, cache lines can be put into drowsy mode more aggressively to save more power."[2]

4. Sub-threshold logic

The 2002 drowsy cache paper cites advances in short-channel effects: "Due to short-channel effects in deep-submicron processes, leakage current reduces significantly with voltage scaling. The combined effect of reduced leakage current and voltage yields a dramatic reduction in leakage power."[3] A 1990 paper developed the new model for short-channel effects and their benefit to sub-threshold logic.[30] A key advance was that the alpha power law model was reduced from $\alpha = 2$ to $\alpha = 1.3$. This was implemented in the calculation of the switching speed:[4d]

$$\text{Delay} \; \alpha \; 1/I_{on} \; \alpha \; V_{dd}/(V_{dd} - V_T)^a$$

Martin et al applied the exponent, a, of the alpha power delay model of an inverter, with a=1, to the dynamic power law to produce the formula for performance, *f*: [4e]

$$f = (L_d K_6)^{-1} ((1 + K_1) V_{dd} + K_2 V_{bs} - V_{th1})^a$$

The combined formula of DVS and Adaptive Body Biasing (ABB), led to a further 48% reduction in energy over DVS (drowsy logic) alone six months later.[31] Other labs have achieved similar results.[32,33] Today, an 85% reduction in leakage power can be found in IoT devices which are designed for batteryless operation and energy-harvesting. Microcontrollers by companies such as Ambiq Micro are known to achieve a 13-fold reduction compared to other chips.[34] IoT device-makers such as ONiO feature an integrated solar/RF/thermoelectric harvester and power management integrated circuit with low-power, asynchronous ROM/RAM.[35] Since the latter uses just 2KB of RAM, it may not need to operate in sub-threshold voltage as much as Ambiq's 2MB of MRAM, which can still achieve ultra low energy consumption.[36] Furthermore, the ultra low power of sub-threshold voltage combined with current drowsy cache techniques suggests mobile phones could one day run on solar power.

5. SRAM and in-memory computing

While drowsy modes have been developed for both instruction and data cache, in high-performance computing (HPC), it may not yet have some of the advantages of state-of-the art memory such as spin-transfer torque (STT) MRAM and spin-orbital torque (SOT) MRAM.[37] That is because they are designed to operate in low-data modes where speed is not as crucial to operation such as remote sensors with fixed interval telemetry. That said, the fast-wake up is known to be 1-2 cycles.[2]

6. Conclusion: Amdahl's & Landauer's Limits

The 2002 paper cites Amdahl's Law in calculating the theoretical minimum.[3] While it does acknowledge advances in short-channel effects and subsequent research detailed new avenues for advancing Moore's Law,[38] Amdahl's Law is more relevant to parallel multi-core architectures, of which sub-threshold voltage microprocessors do not necessarily adopt: "The lowest supply voltage at which a logic gate can operate while still acting as an amplifier is only a few times larger than k_BT/q."[39] As Scott Hanson recently stated that "Moore's law is alive and well for the embedded world. We're at a process node today that's 22 nanometers," one can speculate on the performance gains yet to come.[36]

References

[1] https://semiengineering.com/a-power-first-approach/

[2] "Drowsy Caches: Simple Techniques for Reducing Leakage Energy—A Retrospective" Krisztián Flautner, Nam Sung Kim, Steve Martin, David Blaauw, Trevor Mudge; ISCA@50 25-Year Retrospective: 1996-2020

https://bpb-us-w2.wpmucdn.com/sites.coecis.cornell.edu/dist/7/587/files/2023/07/drowsy_retro.pdf

[3] "Drowsy Caches: Simple Techniques for Reducing Leakage Power" Krisztián Flautner, Nam Sung Kim, Steve Martin, David Blaauw, Trevor Mudge (2002) https://web.eecs.umich.edu/~manowar/publications/drowsy-caches-ISCA2002.pdf

[4] *Computer Architecture Techniques for Power-Efficiency*, Stefanos Kaxiras & Margaret Martonosi ([a]Ch. 5.3)([b]Ch. 3.1)([c]Ch. 1.2.1)([d]Ch. 5.1.1)([e]Ch. 5.4.1)([f]Ch.4.5.1) Morgan & Claypool Press 2008

[5] "Gated-Vdd: A Circuit Technique to Reduce Leakage in Deep-Submicron Cache Memories," 2000 Michael Powell, Se-Hyun

Yang, Babak Falsafi, Kaushik Roy, and T. N. Vijaykumar https://engineering.purdue.edu/~vijay/papers/2000/gatedvdd.pdf

[6] "Cache Decay: Exploiting Generational Behavior to Reduce Cache Leakage Power" 2001 Stefanos Kaxiras, Zhigang Hu, Margaret Martonosi https://mrmgroup.cs.princeton.edu/papers/hzg-isca2001.pdf

[7] "Adaptive Mode Control:A Static-Power-Efficient Cache Design" Huiyang Zhou, Mark C. Toburen, Eric Rotenberg, Thomas M. Conte 2001 https://prod.tinker.cc.gatech.edu/journal/zhou03adaptive.pdf https://pdfs.semanticscholar.org/ba53/2536b7d1890f967d1aac7485544f6abebe39.pdf

[8] "RETROSPECTIVE: Cache Decay: Exploiting Generational Behavior to Reduce Cache Leakage Power" 06/2023 https://bpb-us-w2.wpmucdn.com/sites.coecis.cornell.edu/dist/7/587/files/2023/06/Hu_2001_Cache.pdf

[9] "BTB Access Filtering: A Low Energy and High Performance Design" Shuai Wang, Jie Hu, and Sotirios G. Ziavras Department of Electrical and Computer Engineering New Jersey Institute of Technology (2008) https://web.archive.org/web/20090920084956id_/http://web.njit.edu:80/~sw63/pub/ISVLSI_BAF_2008.pdf

[10] https://en.wikipedia.org/wiki/List_of_ARM_processors

[11]https://web.archive.org/web/20130926155924/http://www.eetimes.com/document.asp?doc_id=1208831 "VLSI Technology Now Shipping ARM810" 08/1996

[12]https://web.archive.org/web/20181224080542/https://www.hotchips.org/wp-content/uploads/hc_archives/hc08/2_Mon/HC8.S4/HC8.4.1.pdf "ARM810: Dancing to the Beat of a Different Drum" 07/1996

[13] https://en.wikipedia.org/wiki/%CE%9CClinux

[14] https://www.emcraft.com/

[15] https://developer.arm.com/documentation/dai0298/a

[16]https://docs.zephyrproject.org/latest/hardware/arch/arm_cortex_m.html

[17] https://docs.zephyrproject.org/latest/kernel/services/other/float.html

[18] "Deterministic Clock Gating for Microprocessor Power Reduction" 2003 https://engineering.purdue.edu/~vijay/papers/2003/dcg.pdf

[19] "Software-Defined Hardware Architectures" 05/2023 https://semiengineering.com/software-defined-hardware-architectures/

[20] "Why Chiplets Don't Work For All Designs" 09/2023 https://semiengineering.com/why-chiplets-dont-work-for-all-designs/

[21] "Is UCIe Really Universal?" Arteris IP, 11/2022 https://semiengineering.com/is-ucie-really-universal/

[22] https://pulp-platform.org/community/showthread.php?tid=229&pid=650#pid650 forum post, 2021

[23] "S-L Power Modeling Gains Steam" 08/2014 https://semiengineering.com/system-level-power-modeling-activities-get-rolling/

[24] "With Responsibility Comes Power," 02/2015 https://semiengineering.com/with-responsibility-comes-power/

[25] "Hybrid Drowsy SRAM and STT-RAM Buffer Designs for Dark-Silicon-Aware NoC", 2016 Jia Zhan, Student Member, IEEE, Jin Ouyang, Member, IEEE, Fen Ge, Member, IEEE, Jishen Zhao, Member, IEEE, and Yuan Xie, Fellow, IEEE https://cseweb.ucsd.edu/~jzhao/files/darlsilicon-noc-tvlsi2016.pdf

[26] "Design the efficient SRAM circuit using 4transistor with sleepy logic" International Journal of Pure and Applied Mathematics Volume 118 No. 20 2018, 115-123 https://acadpubl.eu/hub/2018-118-21/articles/21b/15.pdf

[27] "Design and Implementation of Low Power SRAM Using Highly Effective Lever Shifters", 2021 https://scholar.uwindsor.ca/cgi/viewcontent.cgi?article=9892&context=etd

[28] "Low Power CMOS Design" A.P. Chandrakasan; S. Sheng; R.W. Brodersen, 1992 https://mtlsites.mit.edu/researchgroups/icsystems/pubs/journals/1992_chandrakasan_jssc.pdf

[29] "Design and Optimization of Low Voltage High Performance Dual Threshold CMOS Circuits" Liqiong W ei, Zhanping Chen, Mark Johnson, Kaushik Roy & Vivek De, 1998 https://dl.acm.org/doi/pdf/10.1145/277044.277179

[30] "A JSSC Classic Paper: The Simple Model of CMOS Drain Current" 10/2004 https://www.eng.auburn.edu/~agrawvd/COURSE/READING/LOWP/alpha_power_law.pdf

[31] "Combined Dynamic Voltage Scaling and Adaptive Body Biasing for Lower Power Microprocessors under Dynamic Workloads", 11/2002 https://ieeexplore.ieee.org/document/1167611

[32] "Combined Dynamic Voltage Scaling and Adaptive Body Biasing for Heterogeneous Distributed Real-time Embedded Systems" Le Yan, Jiong Luo and Niraj K. Jha ICCAD'03, November 11-13, 2003 https://web.archive.org/web/20050223092839id_/http://www.princeton.edu:80/~lyan/pub/iccad03.pdf

[33] "Impact of process scaling on the efficacy of leakage reduction schemes" 2004 Yuh-Fang Tsai, David Duarte, N. Vijaykrishnan, Mary Jane Irwin https://www.micromagic.com/news/icicdt04final.pdf

[34] "What's All This Subthreshold Stuff, Anyhow?" 02/2019 https://www.electronicdesign.com/technologies/analog/article/21807652/whats-all-this-subthreshold-stuff-anyhow

[35] "What if You Never Had to Charge Your Gadgets Again?" 01/2024 https://www.wsj.com/tech/personal-tech/what-if-you-never-had-to-charge-your-gadgets-again-955ea960

[36] "Interview With Scott Hanson - Founder and CTO at Ambiq" 01/2024 https://www.safetydetectives.com/blog/scott-hanson-ambiq/

[37] "TSMC tandem builds exotic new MRAM-based memory with radically lower latency and power consumption" 01/2024 https://www.tomshardware.com/pc-components/dram/tsmc-tandem-builds-exotic-new-memory-with-radically-lower-latency-and-power-consumption-mram-based-memory-can-also-conduct-its-own-compute-operations

[38] "Near-Threshold Computing: Reclaiming Moore's Law Through Energy Efficient Integrated Circuits" January 2010 https://ieeexplore.ieee.org/document/5395763

[39]*Recent Progress in Boolean Logic*, Bernd Steinbach, Vincent C. Gaudet, 2013 (4.1.3, 204),(4.1.5, 211)